This CVPR Findings paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

SCAIL: Towards Studio-Grade Character Animation via In-Context Learning of 3D-Consistent Pose Representations

Wenhao Yan[1*†] Sheng Ye[1*†] Zhuoyi Yang[1†‡] Jiayan Teng[1†] ZhenHui Dong[1]
Kairui Wen[1] Xiaotao Gu[2] Yong-Jin Liu[1§] Jie Tang[1§]
[1]Tsinghua University [2]Z.ai



Figure 1. We propose **SCAIL**, a character animation framework that enables high-fidelity character animation under diverse and challenging conditions, including large motion variations, stylized characters, and multi-character interactions.

Abstract

*Achieving controllable character animation that meets studio-grade standards remains challenging despite recent progress. Existing approaches can transfer motion from a driving video to a reference image, but often fail to preserve structural fidelity and temporal consistency in wild scenarios involving complex motion and cross-identity animations. In this work, we present **SCAIL** (a framework toward **S**tudio-grade **C**haracter **A**nimation via **I**n-context **L**earning), which is designed to address these challenges from two key innovations. First, we propose a novel 3D pose representation, providing a robust and flexible motion signal. Second, we introduce a full-context pose injection mechanism within a diffusion-transformer, enabling effective spatio-temporal reasoning over full motion sequences. To align with studio-grade requirements, we develop a curated data pipeline ensuring both diversity and quality, and establish a comprehensive benchmark for systematic evaluation. Experiments show that **SCAIL** achieves state-of-the-art performance and advances character animation toward studio-grade controlling. Code and model are available at zai-org/SCAIL.*

1. Introduction

High-fidelity character animation has tremendous potential for film production. Conventional filmmaking pipelines rely on a complicated workflow—involving motion capture, rigging, rendering—that demands expensive hardware and significant expert labor. Recently, the emergence of video generation models [2, 14, 37, 48] introduces a new paradigm: given a reference image and a driving video, such models can directly synthesize animations that follow the target motion in driving video. Similar to real-world production pipelines, existing methods [5, 16, 31, 32, 39, 45, 50, 54] typically begin by extracting skeletal motion sequences from the driving video as a form of "motion capture", and then inject this information into a video generation model to perform "rigging" and "rendering". However, these methods often struggle with challenging scenarios, such as com-

*Equal contribution.
†Work done during internship at Z.ai.
‡Project leader.
§Corresponding author: jietang@tsinghua.edu.cn, liuyongjin@tsinghua.edu.cn

plex motions (*e.g.,* turning, rolling, flipping), multi-person interactions (*e.g.,* dancing, hugging, fighting), and cross-domain animation where the reference and driving subjects differ significantly in appearance or body shape, exhibiting distorted appearance or implausible limb occlusions. In this work, we identify these challenges as the primary bottleneck to achieving studio-grade controllable animation. These limitations reveal that current skeletal pose representations and pose-driving generation methods fail to adequately capture 3D structure of driving motions, inter-character spatial and occlusion relationships, and temporal correlations of motion sequences.

First, for motion representation, prior works [5, 16, 18, 32, 39, 50] typically rely on 2D keypoints (extracted from DWPose [47], ViTPose [44], *etc.*), which suffer from noisy predictions and cannot encode occlusion. Some works [38, 54] adopt SMPL [21] mesh controls, which offer strong 3D human priors but cause severe identity leakage. Second, regarding pose injection, a common approach in diffusion transformer (DiT) [24] model is to concatenate conditions channel-wise [5, 53]. However, we find that this method provides more local motion cues than capturing global motion dependencies.

To address the above limitations, we present **SCAIL**, a framework that revisits two core technical bottlenecks: (1) how to build a pose representation that effectively bridges driving and generated videos with **unambiguous and accurate motion encoding**, and (2) how to inject pose control in a way that enables the model to **capture spatio-temporal motion structures**. We propose a novel 3D pose representation that rasterizes bones as spatial cylinders into the pixel plane, such representation can be augmented and retargeted to enable seamless controlling across diverse characters and scenarios. To prevent model from following local motion cues as a shortcut, we propose a full-context pose injection mechanism within a DiT [24]-based architecture. This design allows the model to attend to the entire pose sequence when generating each frame, enabling it to reason about motion context across time, and better capture high-level motion semantics.

To obtain reliable video-pose pairs, we apply a segmentation-and-extraction approach to get 3D keypoints in human interactions. Building on that, a data curation pipeline ensuring both character diversity and motion complexity is established, with automatic filtering based on human presence and motion-based metrics and a final manual stage to select a finetuning set of superior quality. We also observe that current evaluations lack a comprehensive benchmark that adequately reflects production-level requirements. To address this gap, we propose **Studio-Bench**. It consists of two parts: the first evaluates motion adherence and structural integrity under complex single- and multi-person actions, while the second measures model generalization when the reference image and driving video differ in identity or domain. Our **Studio-Bench** covers challenging real-world cases and provides a realistic and rigorous measure of studio-level generation capability.

Our main contributions can be summarized as follows: (1) We propose an identity-agnostic 3D pose representation that serves as motion-driving signal suitable for complex motions and multi-human interactions. (2) We inject driving-pose controls via in-context reasoning to enable effective spatiotemporal motion modeling, yielding superior results in complex and multi-person scenarios. (3) We construct a pipeline for curating high-quality, diverse training data, and establish a comprehensive **Studio-Bench** for systematic evaluation. (4) Our **SCAIL** framework achieves state-of-the-art performance over existing baselines, and advances character image animation toward production-level readiness.

2. Related Work

2.1. Diffusion Models for Video Generation

Diffusion models [12, 29] effectively overcome the training instability and mode collapse of generative adversarial networks (GANs) [9], and emerge as the dominant paradigm in visual content generation. The success of the Stable Diffusion (SD) [25, 26] in image synthesis has naturally inspired the extension from image to video. Subsequently, the Diffusion Transformer (DiT) [24] architecture, combined with RoPE [30] for position encoding, offers superior modeling capability and scalability, becoming the leading backbone for high-quality video generation [19, 23, 37, 48]. More recently, diffusion-based video generation has advanced toward controllability, enabling control over camera viewpoints [10, 43], motion trajectories [7, 51], scene structures [6, 20, 42], subject reference [18, 52], and holistic completion [3]. In this work, we focus on character video generation with precise pose control.

2.2. Controllable Character Image Animation

Character image animation aims to generate photorealistic and temporally coherent video, where the appearance remains consistent with a given reference image and the motion follows a driving video. AnimateAnyone [16] extracts 2D skeletons from the driving video as motion guidance, and designs Pose Guider and ReferenceNet modules to control motion and appearance. Subsequent works explore various improvements. Champ [54] integrates multiple motion signals, including depth maps, SMPL [21] normals, and 2D skeletons. Animate-X [31] introduces a skeleton augmentation strategy to enable animation of anthropomorphic characters with large body-ratio discrepancies. Moreover, MimicMotion [50] address hand and facial distortions via regional loss. DanceTogether [4] extends this line of

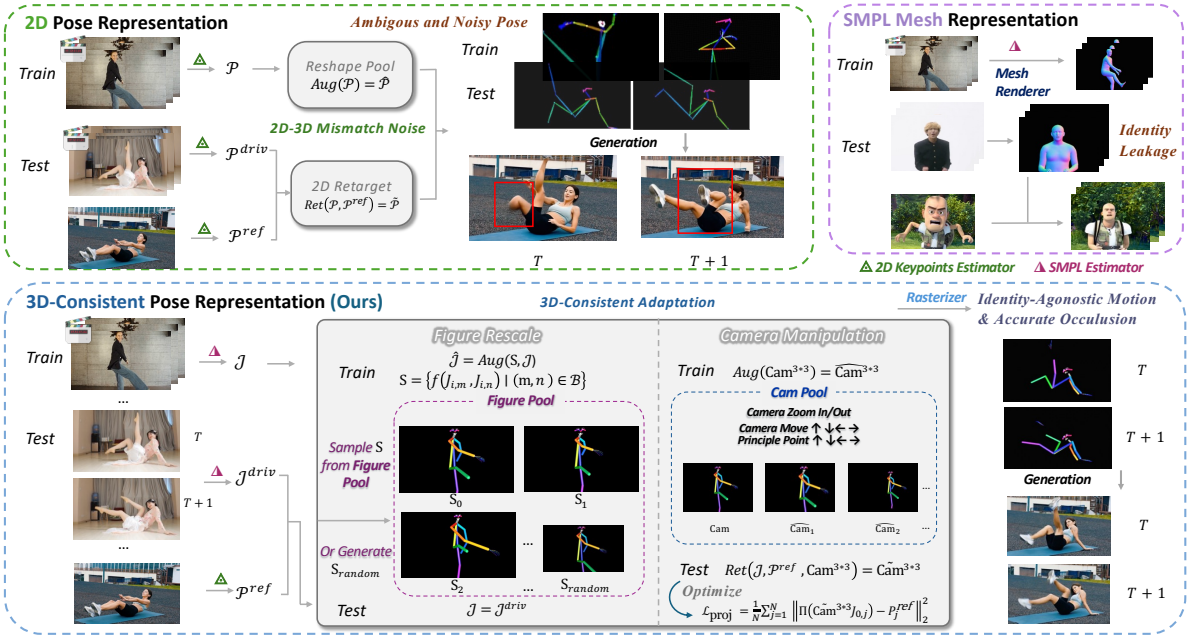


Figure 2. Overview of the proposed 3D-consistent pose. For scaling implementation, we take the clavicle or the pelvis as the central reference, applying scaling from proximal to distal along each limb in bones set $\mathcal{B}$. $Aug(\cdot)$ denotes augmentation in training, $Ret(\cdot)$ denotes retargeting in inference, and $\mathcal{P}^{\text{ref}} = \{\mathbf{P}_j^{\text{ref}} \mid 1 \le j \le N\}$ denotes N estimated 2D keypoints in the reference image. We further incorporate hand and face controls by overlaying 2D hand and face keypoints onto the rendered sequences, and align them with the projection of 3D joints during augmentation or retargeting. For better clarity, we omit the drawing process of 2D hand and face in the figure.

research to multi-person animation. UniAnimate-DiT [39], VACE [18] and Wan-Animate [5] replace U-Net with Transformer as the denoising backbone, significantly improving the generation quality. Despite progress, current models still struggle with complex scenarios and cross-pair animations, which hinders their deployment in real production.

3. Method

3.1. Preliminaries

Latent Diffusion Models. Latent diffusion models [1, 26] reduce the computational cost of pixel-space diffusion [12, 13] by operating in a compressed latent space. Given an input video $\boldsymbol{x}$, a pretrained VAE encoder [37] $\mathcal{E}$ first maps it into a latent representation $\boldsymbol{z}_0 = \mathcal{E}(\boldsymbol{x})$. During the forward diffusion process, Gaussian noise is progressively added over T timesteps, formulated as:

$$q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) = \mathcal{N}\big(\boldsymbol{z}_t; \sqrt{1-\beta_t}\,\boldsymbol{z}_{t-1},\, \beta_t\mathbf{I}\big), \tag{1}$$

where β_t denotes the noise schedule. The denoising model $\boldsymbol{\varepsilon}_\theta(\boldsymbol{z}_t, t, c)$ learns to predict the added noise conditioned on optional input c (*e.g.,* pose, text), and is optimized by:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{z}_t,\, \boldsymbol{\varepsilon}\sim\mathcal{N}(0,\mathbf{I})}\big[\|\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_\theta(\boldsymbol{z}_t, t, c)\|_2^2\big]. \tag{2}$$

Diffusion Transformer (DiT). Traditional diffusion models typically rely on U-Net as denoising backbones. Recent works propose Diffusion Transformer (DiT) [24] architecture, which supports variable input resolution and sequence length. The DiT combines the generative power of diffusion models with the flexibility of Transformers [35], enabling better scalability and stronger modeling capacity.

3.2. 3D-Consistent Pose Conditioning

3D-Consistent Pose Representation. Recent methods [5, 16, 18, 31, 39] that employ 2D keypoint-based skeletons as motion controls perform well on simple actions. However, they often struggle under complex studio-level production scenarios, leading to abnormal human kinematics. As there exists discrepancy between the reference image and the driving pose (e.g. body shape variations), previous 2D-based methods [31, 39] mitigate such gap through 2D skeleton scaling during training or heuristic retargeting at test time. However, such adaptation inherently suffers from deformation due to inconsistency with 3D motion dynamics, further amplifying estimation noise and inaccuracies.

We design our 3D pose representation to retain depth and occlusion, deliver accurate motion and remain identity-agnostic. To achieve that, we employ NLFPose [27] to estimate 3D body keypoints $\mathcal{J} = \{\mathrm{J}_{i,j}\}_{i=1,j=1}^{T,N}$, where $\tilde{\mathrm{J}}_{i,j}$ denotes the predicted 3D joint j at frame i, and connect them according to the skeletal topology in 3D space. We discard SMPL mesh-like representations, which are inherently human-specific and offer limited controllability for

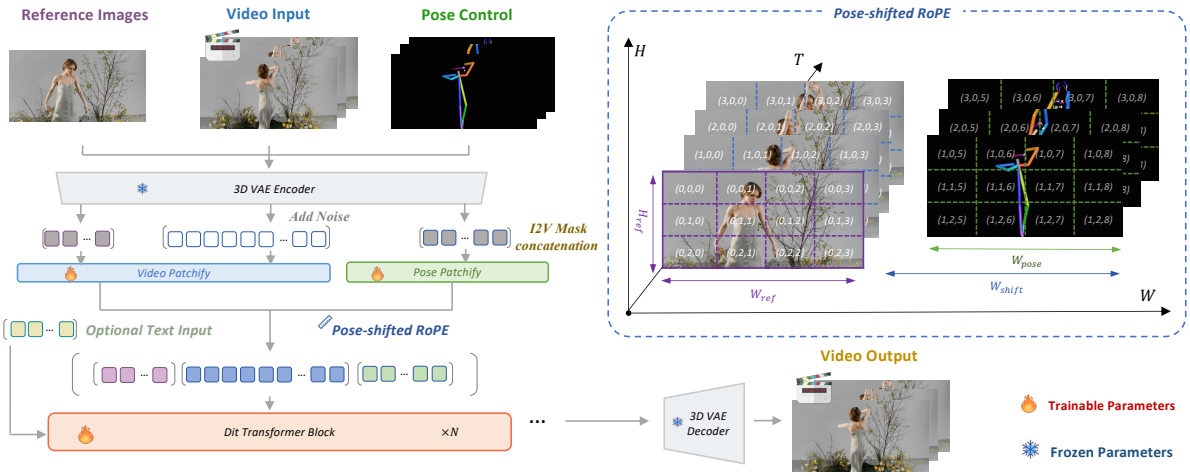


Figure 3. Overview of **SCAIL**'s model architecture. **SCAIL** builds upon I2V model and incorporate pose control as an explicit context for the model to learn spatial-temporal motion. To accommodate to the training setting where reference image and video input are sampled from different parts of the video, we modify the I2V model's input structure by concatenating the reference image at the beginning of the sequence and initiating generation from $T = 1$, using the original I2V pattern to inject the reference CLIP feature. To help the model better distinguish the conditional tokens and the noisy video sequence, we leverage the original mask mechanisim of Wan-I2V model architecture, applying an all-one mask for the reference image and the driving sequence, and an all-zero mask for the noisy video sequence.

augmentation or retargeting, and render 3D skeletons as cylindrical segments to provide driving signals. We distinguish person identities by directly rendering skeletons with different hues instead of relying on hardcoded network designs [4].

To address the disparity between the reference image and the driving pose, we introduce a two-stage optimization strategy leveraging the 3D prior of our representation. Unlike previous methods [31] that applies 2D scaling in training augmentation and non-robust skeleton scaling rules in inference retargeting, we design training pose to be fully identity agnostic with accurate bone angles, and in inference we avoid altering the original estimated skeleton and only optimize the camera instead. To achieve identity decoupling during training, we synthesize a set of scale parameters $\mathcal{S}$, and randomly sample $S_i \in \mathcal{S}$ during training, each S_i denoting a set of scaling parameter of each bones, simulating binding motion to characters with diverse body shapes. Since the underlying estimated skeleton topology and relative bone directions remain consistent, the scaled skeletons is temporally coherent and still faithfully represent the same motions.

In inference, we perform spatial alignment using a projection loss ($\mathcal{L}_{\text{proj}}$) to remap the predicted 3D driving poses to the reference 2D frame to help the model anchor where the motion should be transferred to, and optimize the training process to Specifically, we optimize the following objective to refine the camera projection matrix $\tilde{\text{Cam}}^{3\times3}$:

$$\mathcal{L}_{\text{proj}} = \frac{1}{N}\sum_{j=1}^{N}\left\|\Pi\left(\tilde{\text{Cam}}^{3\times3}\tilde{\text{J}}_{0,j}\right) - \text{P}_j^{\text{ref}}\right\|_2^2, \qquad (3)$$

where $\Pi(\cdot)$ is the perspective projection, N is joint number, and P_j^{ref} is the estimated 2D joints of the reference frame.

As the alignment stage in inference may introduces approximation error, we simulates such error in training by directly applying such aligning method using the reference image and the first frame of the driving sequence (the reference image and driving sequence are sampled from different parts of the video) and exerting modest disturbance on camera parameters $\text{Cam}^{3\times3}$ to enhance the model's robustness towards camera variance. Such design strikes a balance between the controllability of location and robust motion transfer.

Full-Context Driving Pose Injection. Our model builds upon a DiT-based Image-to-Video (I2V) model [37] and inject pose control signals for motion guidance. The common practice for pose information injection is through channel concat [5, 53] or using a pose adapter [39], where the driving pose sequence is embedded and added to the noisy video latents. We first implement the channel concat approach as illustrated in Figure 4, following [53]. While this method shows decent pose-following capability in generation, we find it tend to generate unnatural human pose when the motion is complex despite conditioned on accurate and unambiguous pose. A typical example is the turning motion, where the model often fails to correctly distinguish between front and back views, generating awkward postures. We assume that this limitation stems from the per-frame addition scheme, which fails to provide sufficient temporal context for motion understanding. Since motion is inherently sequential, many actions can only be interpreted correctly

within a temporal context. Therefore, we design *full-context pose injection*, which enables the model to reason over the entire pose sequence. Specifically, we directly concatenates conditional pose tokens with noisy video tokens to facilitate spatial-temporal interactions between the two modality. As shown in Figure 4, this strategy handles motion correctly in complex scenarios.

To mitigate the sequence length increase introduced by our context-aware injection scheme, we apply spatial downsampling to the pose video. Empirically, we find that with 2× downsampling the pose following ability is nearly unaffected. Thus, we adopt this as the default setting in both training and testing, achieving balance between generation quality and efficiency.

Shifted RoPE for Pose Context. Conventional pipelines typically extract poses that are spatially aligned with the original video. However, in our setting, augmentations such as scaling and camera transformations often prevent the extracted pose context to be spatially aligned with the driving video. To address this misalignment, we introduce *Pose-Shifted RoPE* mechanism, which enables the model to effectively retrieve driving signals from the augmented full-sequence pose representation.

As illustrated in Figure 3, conventional 3D RoPE encodes position in the form (t, h, w), where t is the temporal frame index and h, w denote spatial dimensions, each starting from zero. In our design, *Pose-Shifted RoPE* applies a shift along the width dimension specifically to pose representations. The positional vectors for each driving pose token d are defined as:

$$\text{Pos} = [t,\ h,\ W_{\max} : W_{\max} + \text{shift}_W], \tag{4}$$

where shift_W is a constant shift magnitude and $W_{\max}$ is the width of the reference image. As show in Figure 3, this design helps the model to distinguish driving pose tokens from reference image tokens and noisy video tokens in conjunction with the modified I2V Mask and reference image token injection strategy, enhancing the overall performance. Considering the modality gap between the noisy video tokens and the driving pose tokens, we observe that the model performs best particularly when the shift distance is a relatively large constant. To accommodate different downsampling ratios for pose context, mean-pooling on the 3D-RoPE [30] frequencies is performed according to the applied ratio.

3.3. Data Pipeline

Pose Extraction. We adopt [27] for 3D keypoints extraction. However, the default extraction pipeline often fails to detect correct limbs under occlusion, especially in multi-human interactions. To address so, we apply a segmentation-and-extraction approach to obtain reliable

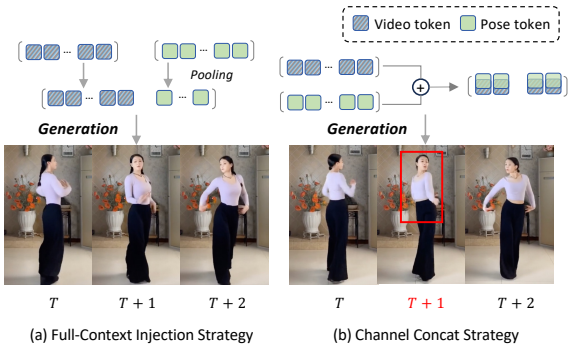


Figure 4. Exploration of different strategies for pose injection.

poses in human interactions. To be more specific, we employ Samurai [46] to track and split the mask of each character, generating multiple single-human video splits. Subsequently, 3D keypoints are extracted for each video that contain one main character. Finally, The 3D skeletons are represented as cylindrical segments in 3D space. Those cylindrical segments are then composed together in 3D space and rasterized on the 2D canvas. This procedure preserves inter-person occlusion relationships, benefiting from NLFPose's accurate depth estimation. In practice, we find that our proposed multi-stage pipeline provides more accurate estimation than direct multi-person pose estimation methods like PromptHMR [40], especially in wild cases involving complex interaction. We follow the same extraction pipeline in inference.

Data Curation. To meet the requirements for generating studio-grade character animation, we have established the following criteria for data filtering: (1) the character should be the primary focus of each frame, (2) the person should exhibit explicit motion, and (3) the pose should be visually complete, covering either the upper body or the full body. We first detect human presence using YOLO [34] and discard clips where characters are not the main subjects. Next, 2D keypoints are extracted using DWPose [47], retaining only half-body or full-body videos. After this stage, multi-person videos can be further separated based on the number of bounding boxes detected by YOLO [34]. We further employ a large language model [8] to analyze captions, complementing the complex rule-based logic of multi-person motion bounding box checking. To compel the model to learn from complex motion dependencies, we compute the *motion speed* from the estimated 3D keypoints and discard samples with limited motion, resulting in a clean and motion-rich training set. The *motion speed* v is calculated as:

$$v = \frac{1}{T}\sum_{t=1}^{T-1}\sum_{j}\left\|\bar{\mathrm{J}}_{t+1,j} - \bar{\mathrm{J}}_{t,j}\right\|_2 \tag{5}$$

where j refers to the joints in the screen and $\bar{\mathrm{J}}_t = \mathrm{J}_t - \mathrm{J}_t^{root}$ denotes the 3D human joint positions relative to the body

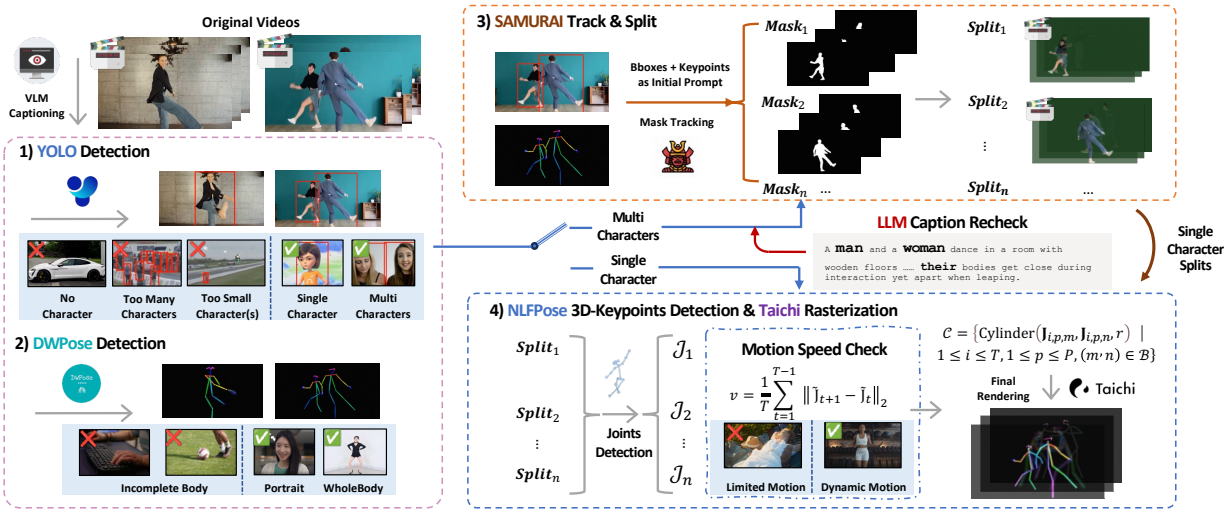


Figure 5. The data curation pipeline. We perform character filtering and motion-speed filtering to construct high-quality training data.

center at frame t.

We collect around 250K high-quality motion-rich video-pose pairs using the filtering pipeline, among which 20K pairs featuring multiple characters. To further enhance model's performance on complex motions, we select 12K samples with the highest *motion speed* from dance, general motion, and obtain a finetuning set of 4,000 high-dynamic videos with minimal blur after manual clarity inspection.

3.4. Studio-Bench for Evaluation

To evaluate the model capabilities under studio-grade standards, we propose **Studio-Bench**, a new evaluation benchmark tailored for Studio-Grade character animation. Previous evaluations [16, 32, 39, 45, 50, 54] primarily focus on simple actions, which fail to capture the challenges present in film production, such as complex and dynamic motions, multi-person interactions, and cross-domain animations. Specifically, our benchmark consists of two parts: *Self-Driven* and *Cross-Driven*. The first part evaluates motion adherence under complex actions. We examine whether the generated videos maintain correct body structure during large-scale motions, and whether the model can accurately capture inter-person relationships in multi-character scenes, totally containing 130 clips. Quantitative metrics can be computed by directly comparing generated results with paired ground-truth videos for this subset. The second part measures the model's transferability when the reference image and driving video differ in identity or domain. We manually select a set of real reference images to test generalization, and further construct additional reference images using an image generation tool [28] to introduce variations in style and body shape, obtaining 120 single-character pairs and 10 multi-character pairs. As we simulate in-the-wild reference-driving discrepancy, pose retarget should be enabled for single-character pairs in this subset. We incorporate the retarget logic of Unianimate-dit [39] for VACE [18] which do not natively support retargeting. For this subset, we conduct user study to assess the generation quality. More details are provided in Suppl. A4.

4. Experiments

4.1. Implementation Details

We train two versions (1.3B and 14B) of our model. The 1.3B model is finetuned from the Wan2.1-1.3B-Fun-Inp backbone on our pretraining dataset for 6,000 steps with a batch size of 96 and a learning rate of 1e-5, using 32 NVIDIA H100 GPUs for approximately two days. For the larger 14B model, we finetune it from the Wan2.1-I2V-14B backbone in two stages: during the pretraining stage, we train for 8,000 steps with a batch size of 96 at a learning rate of 1e-5; after convergence, we perform an additional finetuning stage for 400 steps with the same batch size and a reduced learning rate of 4e-6. Training of the 14B model is conducted on 128 NVIDIA H100 GPUs for over four days with sequence parallelism enabled. All models are optimized using AdamW [22]. During inference, we set the classifier-free guidance (CFG) scale [11] to 4, offering a favorable balance between pose following and video fidelity.

4.2. Quantitative Evaluation

We conduct a quantitative comparison with state-of-the-art methods on our proposed **Studio-Bench**. We convey a detailed blinded user study to collect these metrics, letting users vote for the best in baseline comparison. We also conduct a Win/Tie/Lose study to evaluate our performance between two commonly used open-source frameworks, as well as the closed-source commercial product Viggle [36],

Methods	Self-Driven Animation				Cross-Driven Animation			
	PSNR↑	SSIM↑	LPIPS↓	FVD↓	Mot-Acc↑	Kin-Consis↑	Phy-Consis↑	ID-Sim↑
UniAnimate-DiT [39]	17.79	0.637	0.242	362.27	2.5%	1.7%	0.8%	1.7%
VACE [18]	16.73	0.588	0.263	264.71	9.2%	14.2%	18.3%	32.5%
Wan-Animate [5]	18.54	0.648	0.221	187.61	35.0%	28.3%	24.2%	20.0%
SCAIL-14B (Ours)	**19.22**	**0.660**	**0.206**	**176.16**	**53.3%**	**55.8%**	**56.7%**	**45.8%**

Table 1. Quantitative comparison for **SCAIL**-14B and baselines. All compared methods are built upon 14B Wan [37] foundation models.

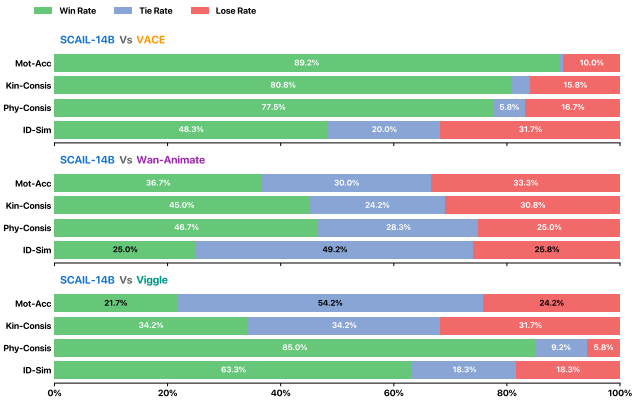


Figure 6. User study for comparing our model with popular community and commercial projects.

omitting UniAnimate-DiT[39] due to obvious artifacts under studio-grade demandings. Viggle is widely believed to rely on a 3D foundation model rather than video diffusion and can be a strong baseline for frame-by-frame motion accuracy under complex scenarios. Results in Table 1 show that our model performs better than current open-source works on **Studio-Bench**. Under the best-model selection scheme, our method shows significant improvements across all metrics, and the win/tie/lose study further highlights its advantages in motion-transfer accuracy over other video diffusion methods and state-of-the-art motion naturalness.

4.3. Qualitative Evaluation

Figure 7 shows qualitative comparisons across both self-driven and cross-driven animation settings. In self-driven cases, **SCAIL** produces animations with stable structure and accurate limb articulation, particularly in challenging motions such as spinning and bending. Baseline methods often exhibit structural distortions or incomplete pose following. In cross-driven cases, our method also demonstrates stronger motion transferability and identity preservation. In contrast, prior methods produce artifacts such as body-shape drift, inconsistent textures, and incorrect interaction geometry. These results highlight **SCAIL**'s ability to handle both large motion variations and cross-domain appearance gaps, yielding videos that are more coherent, natural, and visually appealing.

4.4. Ablation Study

We conduct ablation studies on **SCAIL**-1.3B model to evaluate the contribution of each component. For reference, we also compare against the U-Net based MimicMotion [50], highlighting the advantage of our DiT-based architecture. More ablations would be provided in Suppl. A2.

Ablation on the Pose Representation. To validate the effectiveness of our 3D-consistent pose representation, we compare with a 2D keypoints-based pipeline (extracted using DWPose [47] and denoted as *w/ 2D Pose*). Implementation details of the 2D keypoints-based baseline and qualitative comparisions are also provided in Suppl. A2. Results shows the pose representation is critical for the performance gain, and such improvements are more pronounced under full-context pose injection than under channel-concat scheme.

Ablation on Full-Context Driving Pose Injection. As shown in Table 2, replacing channel concatenation with the proposed full-context injection consistently improves all the metrics when providing the model with temporally coherent and information-rich conditions. Qualitatively, We observe that generated videos exhibit fewer artifacts like limb tearing and structural collapse (shown in Figure 8), indicating that our context-aware injection strategy enables the model to capture global motion dependencies and reason plausible human poses. Specifically, case (2) in 8 shows that through this strategy, when the estimator misidentifies the left part and the right part and extracts a posture like a forward lunge in a running driving video, our model with full-context pose injection can still generate correct running posture based on the global motion, as conditions from the reference frame and estimations in other frames clearly reflect the semantic context of running along the river.

Ablation on Pose-Shifted RoPE. Removing *Pose-Shifted RoPE* (*P-RoPE*) leads to noticeable performance degradation (Table 2), especially in LPIPS and FVD, confirming that precise pose-aware positional encoding is crucial for maintaining image quality and temporal motion smoothness. Figure 8 shows that *P-RoPE* can strengthen the cor-

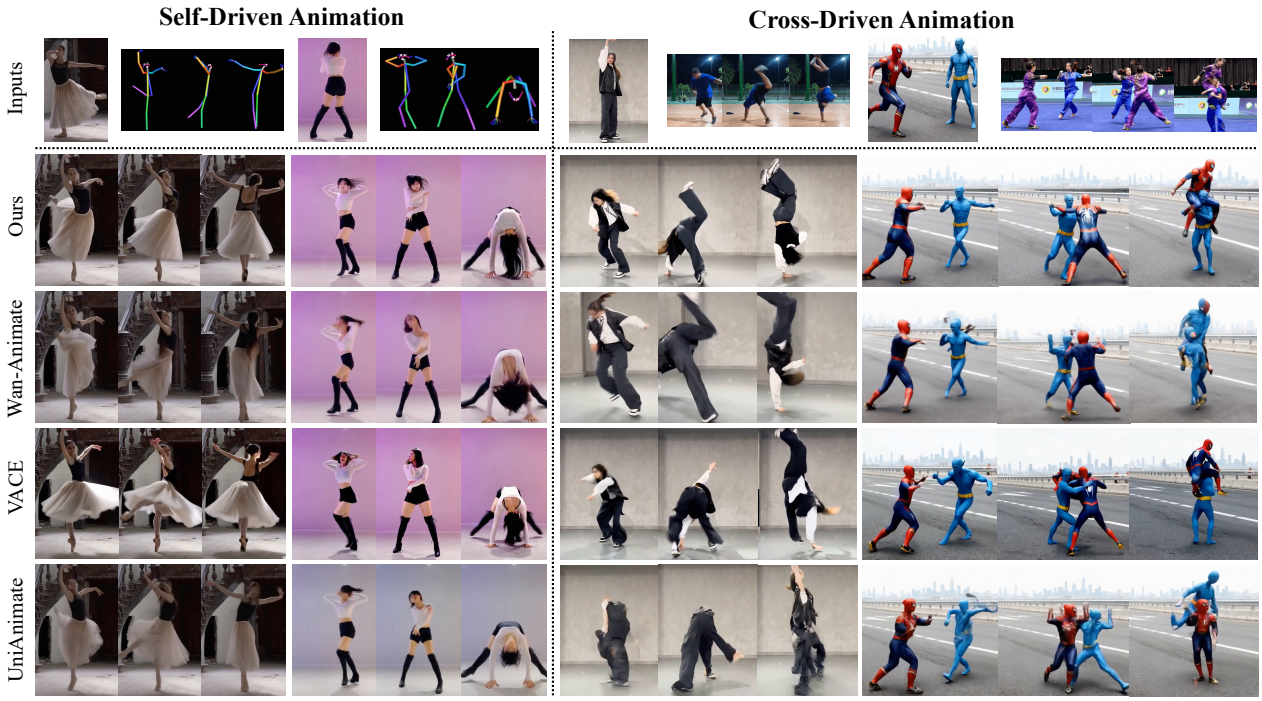


Figure 7. Qualitative comparison for single-character animation. Rendered pose in Cross-Driven Animation are omitted for clarity.

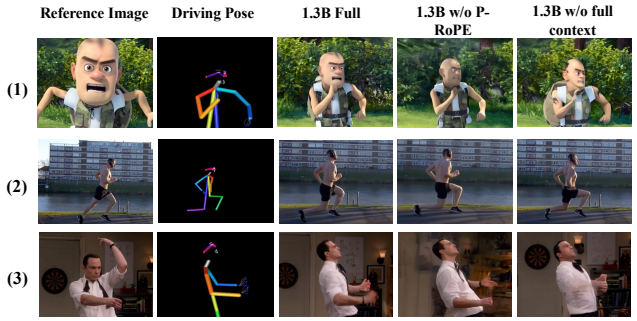


Figure 8. Visualization of ablation study. The *w/o full-context* denotes using the channel-concatenation strategy.

Methods	Self-Driven Animation			
	PSNR↑	SSIM↑	LPIPS↓	FVD↓
SCAIL-1.3B (Ours)	**18.08**	**0.639**	**0.249**	**228.62**
SCAIL-1.3B *w/ 2D Pose*	17.08	0.619	0.284	295.36
SCAIL-1.3B *w/o P-RoPE*	17.79	0.637	0.280	269.35
Channel Concat-1.3B	17.69	0.626	0.262	263.63
Channel Concat-1.3B *w/ 2D Pose*	17.12	0.624	0.282	296.23
MimicMotion [50]	17.01	0.630	0.314	334.24

Table 2. Ablation study conducted on **SCAIL**-1.3B.

respondence between motion cues and spatial structure, resulting in more accurate hand articulation and more reasonable foot grounding. Results from case (1) demonstrate that with full-context injection, *P-RoPE* yields the strongest disentanglement of character identity and pose guidance, enabling the model to faithfully preserve the subject's appearance while following motion patterns.

5. Conclusion

In this work, we present **SCAIL**, a novel framework for studio-grade character image animation. By introducing a scalable and robust 3D pose representation and leveraging a novel full-context pose injection with shifted RoPE, we enhances spatiotemporal reasoning within a DiT architecture, allowing the model to generate structurally accurate and temporally consistent animations under challenging scenarios. With curated data pipeline and comprehensive Studio-Bench, we push character animation towards production-level standards. Experiments show that SCAIL achieves state-of-the-art performance in both self-driven and cross-driven animation. We believe this work provides a solid step toward practical, production-ready character animation.

Acknowledgments

This work was partially supported by the Natural Science Foundation of China (NSFC) 62425601, 62461160309, and NSFC-RGC Joint Research Scheme N_HKU705/24, and the New Cornerstone Science Foundation through the XPLORER PRIZE.

References

[1] Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv preprint arXiv:2304.08477*, 2023. 3

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1

[3] Minghong Cai, Qiulin Wang, Zongli Ye, Wenze Liu, Quande Liu, Weicai Ye, Xintao Wang, Pengfei Wan, Kun Gai, and Xiangyu Yue. Videocanvas: Unified video completion from arbitrary spatiotemporal patches via in-context conditioning. *arXiv preprint arXiv:2510.08555*, 2025. 2

[4] Junhao Chen, Mingjin Chen, Jianjin Xu, Xiang Li, Junting Dong, Mingze Sun, Puhua Jiang, Hongxiang Li, Yuhang Yang, Hao Zhao, et al. Dancetogether! identity-preserving multi-person interactive video generation. *arXiv preprint arXiv:2505.18078*, 2025. 2, 4

[5] Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, et al. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 2025. 1, 2, 3, 4, 7

[6] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7346–7356, 2023. 2

[7] Xiao Fu, Xian Liu, Xintao Wang, Sida Peng, Menghan Xia, Xiaoyu Shi, Ziyang Yuan, Pengfei Wan, Di Zhang, and Dahua Lin. 3dtrajmaster: Mastering 3d trajectory for multi-entity motion in video generation. In *The Thirteenth International Conference on Learning Representations, ICLR*, 2025. 2

[8] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024. 5

[9] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[10] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 2

[11] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3

[13] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 3

[14] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022. 1

[15] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 3

[16] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 1, 2, 3, 6

[17] Yuanming Hu. Taichi: An open-source computer graphics library. *arXiv preprint arXiv:1804.09293*, 2018. 1

[18] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025. 2, 3, 6, 7, 1

[19] Kuaishou AI Team. Kling. https://kling.kuaishou.com/en, 2024. 2

[20] Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. Drivingdiffusion: layout-guided multi-view driving scenarios video generation with latent diffusion model. In *European Conference on Computer Vision*, pages 469–485. Springer, 2024. 2

[21] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, 2015. 2

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR*, 2019. 6

[23] OpenAI. Sora. https://openai.com/index/sora/, 2024. 2

[24] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2, 3

[25] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations, ICLR*, 2024. 2

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3

[27] István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3d human pose and shape estimation. *Advances in Neural Information Processing Systems*, 37: 140032–140065, 2024. 3, 5

[28] Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025. 6

[29] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR*, 2021. 2

[30] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 2, 5

[31] Shuai Tan, Biao Gong, Xiang Wang, Shiwei Zhang, Dandan Zheng, Ruobing Zheng, Kecheng Zheng, Jingdong Chen, and Ming Yang. Animate-x: Universal character image animation with enhanced motion representation. In *The Thirteenth International Conference on Learning Representations, ICLR*, 2025. 1, 2, 3, 4

[32] Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, and Zuxuan Wu. Stableanimator: High-quality identity-preserving human image animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21096–21106, 2025. 1, 2, 6

[33] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 3

[34] Rejin Varghese and M Sambath. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International conference on advances in data engineering and intelligent computing systems (ADICS)*, pages 1–6. IEEE, 2024. 5

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[36] Viggle AI. Viggle. https://viggle.ai/, 2024. 6, 3

[37] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1, 2, 3, 4, 7

[38] Qilin Wang, Zhengkai Jiang, Chengming Xu, Jiangning Zhang, Yabiao Wang, Xinyi Zhang, Yun Cao, Weijian Cao, Chengjie Wang, and Yanwei Fu. Vividpose: Advancing stable video diffusion for realistic human image animation. *arXiv preprint arXiv:2405.18156*, 2024. 2

[39] Xiang Wang, Shiwei Zhang, Longxiang Tang, Yingya Zhang, Changxin Gao, Yuehuan Wang, and Nong Sang. Unianimate-dit: Human image animation with large-scale video diffusion transformer. *arXiv preprint arXiv:2504.11289*, 2025. 1, 2, 3, 4, 6, 7

[40] Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J. Black, and Muhammed Kocabas. Prompthmr: Promptable human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1148–1159, 2025. 5

[41] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 3

[42] Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, et al. Make-your-video: Customized video generation using textual and structural guidance. *IEEE Transactions on Visualization and Computer Graphics*, 31 (2):1526–1541, 2024. 2

[43] Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024. 2

[44] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022. 2

[45] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490, 2024. 1, 6

[46] Cheng-Yen Yang, Hsiang-Wei Huang, Wenhao Chai, Zhongyu Jiang, and Jenq-Neng Hwang. Samurai: Adapting segment anything model for zero-shot visual tracking with motion-aware memory. *arXiv preprint arXiv:2411.11922*, 2024. 5

[47] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 2, 5, 7, 1

[48] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations, ICLR*, 2025. 1, 2

[49] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 3

[50] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. In *International Conference on Machine Learning*, 2025. 1, 2, 6, 7, 8

[51] Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2063–2073, 2025. 2

[52] Zhenxing Zhang, Jiayan Teng, Zhuoyi Yang, Tiankun Cao, Cheng Wang, Xiaotao Gu, Jie Tang, Dan Guo, and Meng Wang. Kaleido: Open-sourced multi-subject reference video generation model. *arXiv preprint arXiv:2510.18573*, 2025. 2

[53] Jingkai Zhou, Yifan Wu, Shikai Li, Min Wei, Chao Fan, Weihua Chen, Wei Jiang, and Fan Wang. Realisdance-dit: Simple yet strong baseline towards controllable character anima-

tion in the wild. *arXiv preprint arXiv:2504.14977*, 2025. 2, 4

[54] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision*, pages 145–162. Springer, 2024. 1, 2, 6